Title

Predicting risk for Alcohol Use Disorder using longitudinal data with multimodal biomarkers and family history: a machine learning study.

Journal

Molecular psychiatry, 26(4)

ISSN

1359-4184

Authors

Kinreich, Sivan
Meyers, Jacquelyn L
Maron-Katz, Adi
et al.

Publication Date

2021-04-01

DOI

10.1038/s41380-019-0534-x

Peer reviewed

Predicting risk for Alcohol Use Disorder using longitudinal data with multimodal biomarkers and family history: A machine learning study

Sivan Kinreich[1], Jacquelyn Meyers[1], Adi Maron-Katz[7], Chella Kamarajan[1], Ashwini Pandey[1], David Chorlian[1], Jian Zhang[1], Gayathri Pandey[1], Stacey Subbie[1], Dan Pitti[1], Andrey Anokhin[3], Lance Bauer[4], Victor Hesselbrock[4], Marc Schuckit[6], Howard J. Edenberg[2,5], Bernice Porjesz[1]

1Department of Psychiatry, State University of New York Downstate Medical Center, Brooklyn, NY, USA
2Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA
3Department of Psychiatry, Washington University School of Medicine in St Louis, St Louis, MO, USA
4Department of Psychiatry, University of Connecticut School of Medicine, Farmington, CT, USA
5Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, USA
6 Department of Psychiatry, University of California, San Diego School of Medicine, La Jolla, CA, USA
7 Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA

Corresponding author:
Sivan Kinreich, Ph.D. SUNY Downstate Medical Center, NY 11203, USA
Tel.: 718-270-2231
E-mail: sivan.kinreich@downstate.edu

Abstract

Predictive models have succeeded in distinguishing between individuals with Alcohol use Disorder (AUD) and controls. However, predictive models identifying who is prone to develop AUD and the biomarkers indicating a predisposition to AUD are still unclear. Our sample (n=656) included offspring and non-offspring of European American (EA) and African American (AA) ancestry from the Collaborative Study of the Genetics of Alcoholism (COGA) who were recruited as early as age 12 and were unaffected at first assessment and reassessed years later as AUD (DSM-5) (n=328) or unaffected (n=328). Machine learning analysis was performed for 220 EEG measures, 149 alcohol-related single nucleotide polymorphisms (SNPs) from a recent large Genome-wide Association Study (GWAS) of alcohol use/misuse and 2 family history (mother DSM-5 AUD and father DSM-5 AUD) features using supervised, Linear Support Vector Machine (SVM) classifier to test which features assessed before developing AUD predict those who go on to develop AUD. Age, gender, and ancestry stratified analyses were performed. Results indicate significant and higher accuracy rates for the AA compared to the EA prediction models and a higher model accuracy trend among females compared to males for both ancestries. Combined EEG and SNP features model outperformed models based on only EEG features or only SNP features for

both EA and AA samples. This multidimensional superiority was confirmed in a follow-up analysis in the AA age groups (12-15, 16-19, 20-30) and EA age group (16-19). In both ancestry samples, the youngest age group achieved higher accuracy score than the two other older age groups. Maternal AUD increased the model's accuracy in both ancestries' samples. Several discriminative EEG measures and SNPs features were identified, including lower posterior gamma, higher slow wave connectivity (delta, theta, alpha), higher frontal gamma ratio, higher beta correlation in the parietal area, and 5 SNPs: rs4780836, rs2605140, rs11690265, rs692854 and rs13380649. Results highlight the significance of sampling uniformity followed by stratified (e.g., ancestry, gender, developmental period) analysis, and wider selection of features, to generate better prediction scores allowing a more accurate estimation of AUD development.

Introduction

Identifying who is vulnerable to develop Alcohol Use Disorder (AUD), determining 'sensitive periods', and finding relevant biological markers are a major challenge. Studies show that rates of alcohol consumption dramatically increase during the teenage years [1] and genetic and environmental factors can increase the risk for transitioning to AUD [2]. However, clear indication as to who is prone to develop AUD is still unclear. Recent studies have suggested that multidimensional modeling of genetic, biological, and psychosocial information may better reflect the underlying pathophysiology compared to one-dimensional measures [3, 4]. Indeed, over the last decade, machine learning (ML) approaches and data mining processes have been successfully applied for analysis of multidimensional datasets including neuroimaging and genetic data to help in the context of disease diagnosis [5, 6], outperforming classical regression approaches [7]. ML Support Vector Machine (SVM) classifiers have succeeded in predicting diagnosis, clinical outcomes, and classifying disorders such as depression [6], schizophrenia [4, 8], and AUD [9-11]. Specifically, AUD classifiers achieved significant accuracy utilizing electrophysiological features such as EEG coherence and spectral power (89.3%)[10, 11], EEG's nonlinear features (91.7%) [9], family history (FH) of AUD and psycho-social features [3, 7], as well as genetic information [3, 12].

However, there are no longitudinal studies that analyzed the predictive model of AUD based on data acquired prior to its development, thus, avoiding the confounding with effects of AUD. Such a model can give important information about biomarkers which can indicate the sensitivity to develop AUD. The current study used longitudinal multidimensional data from COGA (e.g., clinical, electrophysiological, SNP, FH), including individuals of European American (EA) and African American (AA) ancestry. COGA collects data and follows AUD/non-AUD individuals starting as early as age 12, enabling a unique opportunity to compare individual's status before and after AUD developed.

Our central hypothesis was that a multidimensional features model will result in a better prediction than each of the modalities separately (EEG measures and genomic data) and that the addition of a FH feature will further increase the prediction score. In this paper we present a supervised ML method (SVM) to classify individuals before AUD emerged into those who developed AUD years later and those who did not. The analysis incorporates EEG measures, FH information, and data on a set of SNPs derived from recent GWAS of alcohol consumption[12, 13], alcohol dependence[14, 15], and alcohol-related EEG measures[15, 16], as features. An essential aspect of identifying a true classifier is to control for possible effects of confounding variables such as age[17, 18], gender[19, 20], and ancestry[21, 22] which can lead to misclassification of the model [23]. Age, gender and

ancestry stratified analysis can lead to separate, more accurate models for each of the groups.[17, 19, 21] Using stratification to control for the confounding variables, age, gender, and ancestry, we expected to find differences in the prediction models between the groups. We also examined the most discriminative features in the predictive models, enhancing our understanding of brain mechanism/genetics/FH features underlying AUD development, risk and resilience.

Method

Participants

The data comprised 656 participants (376 males and 280 females) from Collaborative Study of the Genetics of Alcoholism (COGA), examined within the age range of 12 to 30 years. Data from six collection sites were included in this study. Ascertainment and assessment procedures of COGA recruits have been described elsewhere [24-26] and in Supplementary Materials. For this study we examined only participants who were unaffected at their first visit and reassessed years later and divided them into two groups: DSM-5 AUD and unaffected controls. The AUD group (n=328, 188 males, 140 females) was defined as those diagnosed as unaffected during the first visit (mean age: 17.88±2.95) and diagnosed with lifetime DSM-5 AUD during a follow-up visit (mean number of

years between visits=7.36±3.01). The control group (n=328, 188 males, 140 females) was age matched (p = 0.5) to the AUD group during the first visit (mean age: 17.69±3.11) and diagnosed as unaffected both at that visit and during a follow-up visit (mean number of years between visits=6.64±3.35) (Figure 1). In a series of analyses, the groups were further divided according to ancestry (EA, AA), age (early adolescence: 12-15 years old, late adolescence: 16-19 years old, and adults: 20-30 years old) [27, 28] and gender (male, female). All groups were matched on age. Ancestry, gender, age and missing values dictated a series of analyses that included different subsets of subjects. Full description of each of the groups can be found in Supplementary Tables 1 & 2.

Procedure

EEG data acquisition and preprocessing

Resting EEG was recorded for four minutes in all participants as they were resting on a comfortable chair in a dimly lit, sound-attenuated RF-shielded booth (Industrial Acoustics, Inc., Bronx, NY, USA). A 64-channel electrode cap (Electro-Cap International, Inc., Eaton, OH, USA) based on the extended 10–20 System [29, 30] was used. Participants were asked to stay awake with eyes closed and not to move. EEG recording and preprocessing procedures are described in Supplementary Materials.

Feature Extraction

EEG extracted features: Full description of the EEG features extraction analysis can be found in Supplementary Materials. The following electrophysiological features were extracted: 1. The spectral power **(40 features)**. 2. Coherence values **(90 features)**. 3. Correlation values **(90 features)**

FH extracted features: Parental AUD data (mother or father DSM-5 AUD) **(2 features).**

Genetic data extracted features:

SNPs (149 features; Supplementary Table 3) were selected based upon associations with EEG and alcohol-related traits from several recent Genome-wide Association Study (GWAS) involving EA and AA populations. These include fast beta EEG[16], alcohol consumption[12, 13], DSM-IV alcohol dependence[14, 15], and maximum number of alcoholic drinks within 24 h[31]. Genotyping, imputation and quality control have been previously reported [32, 33] and can be found in Supplementary Materials.

Feature selection and classification model estimation

Feature selection and model estimation and validation were done separately for every group (i.e. only EEG, only SNPs, combined EEG+SNP, male, female, AA,

EA and different age groups). To control for variables overfitting we used regularization method[4, 34], enhancing the prediction accuracy and interpretability of the statistical model. Specifically, for feature selection we used the least absolute shrinkage and selection operator (LASSO) penalty as described by Tibshirani (1996)[35]. The sparsity property of LASSO (i.e. generating coefficient estimates of exactly zero), makes it attractive for feature selection as it reduces the estimation variance while providing a more interpretable final model [36]. Its application to genomic data [37, 38] has shown that selecting a small number of representative features can achieve satisfactory classification. We first determined the regularization parameter using a 10-fold cross-validation (CV) procedure, with the label: control vs. AUD as the response variable. All features with a non-zero coefficient were retained for subsequent analyses. The reduced set of most discriminant features were fed into the classifier to classify the study participants into their respective groups, i.e., either AUD or controls.

A linear-kernel SVM was trained to distinguish between the two groups in a 10-fold cross-validation (CV) procedure that included parameter optimization. For the 10-fold CV, subjects were randomly divided into ten equal groups, a classifier was then trained on nine of the ten groups and tested on the left-out one. Every fold, the entire dataset was shuffled to insure randomization of the groups. Due to the effect of the random division on the classification results we repeated this process 10

times, averaging the output results. To evaluate model performance, we recorded the number of true positives (TP, number of correctly classified AUD) and true negatives (TN, number of correctly classified controls) scores. Classification accuracy was computed as a ratio of sum of TP and TN divided by the sum of all classified subjects. Area under curve (*AUC*)[3] and F-scores were used to evaluate the classification models, while F was defined by the equation [10, 11, 39]

$$F = (1 + \beta^2) \text{ x } \frac{\text{Precision x recall}}{\beta \text{ x precision} + \text{recall}}$$

and can be interpreted as a weighted harmonic average of precision and recall values [39].The precision is defined as the number of true positives divided by the number of true positives plus the number of false positives and the recall is defined as the number of true positives divided by the number of true positives plus the number of false negatives. Due to an absence of prior information about either precision or recall, the beta value was set to 1. More description of models' calculation and comparison can be found in Supplementary Materials.

Results

Different SVM prediction models with overlapping features and different subsets of subjects divided according to ancestry, age and gender were used for the predictive models. Table 1 summarizes the results of the significant predictive model scores across ancestry, gender and age (see Supplementary Tables 4-6 for

full results), revealing higher scores for the AA than for the EA sample ($p <$ 0.001), for females over males in both EA (borderline trend) and AA ($p_{(EA\ male\ vs.\ female)}$=0.06, $p_{(AA\ male\ vs.\ female)}$=0.03) and for the younger age group over the others in both samples (EA; $F_2 = 76.29$, $p < 0.001$, AA; $F_2 = 8.27$, $p = 0.001$) (Table 1). Table 1 and Figure 2 highlight that the combined model of EEG+SNP was more accurate than the model based on only EEG features or only SNP features for both the AA and EA samples (EA; $p_{(EEG\ vs.\ EEG+SNP)}< 0.001$, $p_{(SNP\ vs.\ EEG+SNP}$ <.001) (AA; $p_{(EEG\ vs.\ EEG+SNP)}< 0.001$, $p_{(SNP\ vs.\ EEG+SNP}$ <0.001). Results were confirmed in a follow up analysis in the AA and EA age groups (AA: $p_{(\text{early adolescence },EEG\ vs.\ EEG+SNP)}$<.001, $p_{(\text{late adolescence s, }EEG\ vs.\ EEG+SNP)}< 0.001$, $p_{(adults,\ EEG\ vs.\ EEG+SNP)}< 0.001$)(EA: $p_{(\text{late adolescence s, }EEG\ vs.\ EEG+SNP)}= 0.002$)(Table 1, Supplementary Figure 1). The EA age groups combined models reached significance in the early and late adolescence age range but did not outperform the EEG based model accuracy (Table 1, Supplementary Figure 1). Gender stratified analyses unveiled higher model accuracy in the AA female group over the male in all three features categories (EEG, SNPs and the combined EEG+SNP model) (whole sample; $p_{(\text{EEG male }vs.\ EEG\ \text{female})}< 0.001$, AA; $p_{(\text{SNP male }vs.\ SNP\ \text{female})}= 0.008$, $p_{(EEG+\text{SNP male }vs.\ EEG+SNP\ \text{female})}< 0.001$) while in the EA group only the combined model EA: $p_{(EEG+\text{SNP male }vs.\ EEG+SNP\ \text{female})}< 0.001$ (Figure 3). Overall, out of all the combined models of EEG+SNP features, the AA & EA female groups achieved the highest accuracy of 79.33% (specificity=71.02%,

sensitivity = 87.67%, *AUC*=0.99, *F*=0.81), 78.91% (specificity=76.82%, sensitivity = 81%, *AUC*=0.9, *F*=0.79), respectively, and the AA & EA early adolescence range age of 79.54% (specificity=79.55%, sensitivity = 79.52%, *AUC*=0.93, *F*=0.79), 74.2% (specificity=68.43%, sensitivity = 79.23%, *AUC*=0.89, *F*=0.76), respectively (full list of the significant models in Table 1). Interestingly, we found gender and ethnicity differences when comparing the addition of the FH feature of mother DSM-5 AUD or father DSM-5 AUD to the combined model of EEG+SNP. For both AA and EA, male and female samples, mother AUD feature increased model accuracy EA:($p_{(son\text{-}mother\ vs.\ EEG+SNP)}<0.001$) ($p_{(daughter\ -mother\ vs.\ EEG+SNP)}=0.02$), AA:($p_{(son\text{-}mother\ vs.\ EEG+SNP)=}0.001$, $p_{(daughter\ -mother\ vs\ EEG+SNP)<}0.001$). Father AUD increased the accuracy of the combined model only for the AA female sample ($p_{(daughter\text{-}father\ vs\ EEG+SNP)\ <}0.001$) (Table 1, Figure 4). Finally, the AA female group with the combined model of EEG+SNP features with the addition of FH of father AUD or mother AUD feature achieved the highest accuracy of 87.55% (father AUD) (specificity =85.71%, sensitivity=89.38%, *AUC*=0.99, *F*=0.89) and 87.11% (mother AUD)(specificity =81.3%, sensitivity=92.92%, *AUC*=0.99, *F*=0.88). Comparing the EA & AA groups' sensitivity and specificity values revealed higher sensitivity values in the AA sample ($p_{(sensitivity)}=0.002$).

Discriminative Features

EEG: Supplementary Table 7 presents a summary of selected shared and group specific features stratified by ancestry and gender for the combined EEG and SNP models. The most consistent EEG predictor shared by all the AUD groups for the combined EEG+SNP based model, which distinguished the participants with AUD from the controls, included lower posterior gamma (e.g., amplitude, coherence, correlation) and higher slow wave connectivity (delta, theta, alpha) in multiple locations (weight ranking for each group/band/frequency in Supplementary Table 7 and Supplementary Tables 8-11). All AUD groups exhibited lower occipital gamma amplitude compared to control (weight ranking 1-4). EA-AUD genders shared lower gamma parietal interhemispheric coherence (male) and amplitude (female) (weight ranking 2,4), AA-AUD genders shared lower delta occipital interhemispheric correlation (weight ranking 1,7) and EA and AA female samples shared lower Frontal-Parietal gamma correlation (weight ranking 2,8). On the other hand, higher theta was revealed in AUD EA male interhemispheric connectivity in the occipital, frontal and temporal lobes (weight ranking 1,4,5) while both female groups showed higher slow wave intrahemispheric connectivity (delta, alpha) in frontal-parietal (AA, EA) (weight ranking 2,8) and temporal-iparietal (EA) (weight ranking 4) lobes. The groups differed in higher frontal gamma ratio and

higher beta correlation in the parietal area (AA male) (weight ranking 2) and lower beta intrahemispheric correlation in the frontal-parietal (EA female) (weight ranking 5).

SNPs: A summary of the most robust SNP predictors for vulnerability to develop AUD is presented in Supplementary Table 12 (the full features ranking is in Supplementary Tables 8-11). EA and AA-AUD females shared one SNP, which was found on chromosome 16 (rs4780836, weight ranking 9,7). Ancestry-gender-specific loci were found for EA AUD female sample on chromosome 17 gene FLII (rs2605140, weight ranking 7), and on chromosome 18 (rs303757, weight ranking 18), and two on chromosome 3 (rs7430178, weight ranking 3), (rs13093097, weight ranking 3).

In the AA female sample, loci were found on chromosome 2 (rs11690265, weight ranking 2), and two on chromosome 18 (rs167336, weight ranking 18), and (rs303754, weight ranking 18) and on chromosome 11 (rs34467936, weight ranking 11), and two on chromosome 16 (rs62057756, weight ranking 16), and (rs28709965, weight ranking 16). A locus was found for EA-AUD male sample on chromosome 19 gene FUT2 (rs692854, weight ranking 19), and for the AA-AUD male sample on chromosome 16 (rs13380649, weight ranking 16). Overall,

females had more SNP features than males ($\#SNP_{(AA\ female)}$ =8, ($\#SNP_{(AA\ male)}$ =1) ($\#SNP_{(EA\ female)}$ =5, ($\#SNP_{(EA\ male)}$ =1) (Supplementary Tables 5-6).

Discussion

Machine learning applications hold promise for creating innovative disease prediction models based on longitudinal data. This study used COGA's rich datasets with EEG, genetic, and FH information acquired from individuals as early as age 12, before developing AUD, and followed years later when they either were diagnosed as DSM-5 AUD or unaffected. This is the first study to formulate a prediction model for those who are predisposed to develop AUD using ML with multidimensional features while considering gender and ancestry. We found higher accuracy rates for the prediction models in AA than EA samples. In both AA and EA samples combining EEG and SNP features resulted in higher accuracy scores than the models based on only EEG features or only SNPs, and these results were confirmed in a follow up analysis (same dataset) within the different AA age groups (early adolescence, late adolescence and adults) and EA late adolescence age group. Gender analyses revealed trend of higher model accuracy in the female group over the male group in both the EA and the AA for all three features categories (EEG, SNPs, and the combined EEG+SNP model). We further found gender differences in model accuracy with parental history of AUD added to the

model. Interestingly, both EA and AA samples showed history of maternal AUD as a discriminative feature, increasing the accuracy of the combined EEG+SNP based model. History of paternal AUD increased the model accuracy over the combined EEG+SNP based model only in the AA females. In both samples, the younger age group achieved higher accuracy score than the two older age groups. Several discriminative EEG and SNP features were identified for each of the models revealing novel gender and ancestry specific AUD predisposition biomarkers. Overall, our findings suggest that higher model accuracy is anchored in a wide range of multidimensional features generated from specific homogenous samples (e.g., gender, age, ancestry). Importantly, identifying group-related specific features will generate formulation of better prediction models.

The ML model based on the combined genetic data and EEG data achieved better classification accuracy than using either alone. These results indicate that these two modalities might reflect somewhat different aspects of AUD etiopathology, and cannot replace each other in terms of portraying the disease, also confirming previous literature results showing the advantage of a ML model using multiple dimensions to classify a disease [3]. Importantly, these results open the door to more personalized approaches to predicting diseases. Models based on different modalities can include features that change over time (i.e., brain

structures and functions)[40] and over human maturation (i.e., behavior and psychology)[41] making it possible to focus on specific groups (such as categorization by age, gender, ancestry, FH, culture, and behavior) to create prediction models where individualization has real value to advance personalized care for patients.

Accurate predictive models rely on an optimal subset of a given feature set for a given population. The given set of features in the current study better predicted AUD females than males, and AA-AUD than EA-AUD, implying the need to continue and search for group-specific variables with importance or 'strength' relatable to each group. For example, the low prediction score of both male groups might relate to the models' limited genetic discriminative features (i.e., only one SNP was implicated in any of the male models) in comparison to the females' models (where 4-5 SNPs were implicated).

Our results indicate that across gender and ancestry, individuals who are vulnerable to AUD have posterior (e.g. occipital, parietal) lower gamma activity. These findings are in accordance with a recent review on the neurophysiological correlates of numerous psychiatric disorders, such as depression, bipolar disorder, anxiety and AUD, showing that the most dominant pattern of change across disorder types is power decreases across higher frequencies[42]. Indeed, we found

lower parietal gamma (amplitude and coherence) in EA in both males and females and lower frontal-parietal gamma connectivity in only female groups in both EA and AA, all together suggesting that lower posterior gamma is not only a disease biomarker but also a predisposition factor for increased vulnerability to develop AUD. Gamma activity has been proposed to promote the feed forward or "bottom-up" flow of information from lower to higher regions of the brain during thalamocortical iterative recurrent activity[43]. Reduction in the gamma band power and connectivity is possibly index disruption in bottom-up communication across the posterior cortex leading to sensory and executive dysfunctions, which may reflect altered cortical integration.

On the other hand, we found increased connectivity of slower wave bands (delta, theta, alpha) for both EA genders. These results confirm previous finding of increased absolute theta log power at all locations on the scalp of eyes-closed EEGs of alcohol-dependent individuals[44] and increased frontal[45] and occipital[46] theta in binge drinkers, as well as, increased interhemispheric[47] and intrahemispheric theta coherence[48] when compared to controls. Increased cortical theta is usually linked to deep resting stages[49], transition to sleep [50] and while practicing meditation [51]. These mental processes relate to the suggested model of the "posterior salience network" unfolded in a functional connectivity analysis

during rest as an interoceptive network, regulating central somatic awareness, physiological reactivity and internal homeostatic states[52, 53]. These results suggest that higher-theta-connectivity alcohol-vulnerable individuals have reduced outside attention over introspect inside attention.

Various SNPs were implicated as salient features in predicting the vulnerability to develop AUD. Interestingly, they varied by gender and ancestry. Moreover, females and males did not share any implicated SNPs, which may shed light on previous discrepancies observed in unstratified studies. One variant on chromosome 16 (rs4780836), previously associated in a large GWAS with alcohol consumption[12], was found both in AA and EA females' models suggesting gender-specificity of this susceptibility marker. While this study focused on individual SNPs from previous GWAS, future studies should aggregate information from a large number of potentially causal SNPs, such as Polygenic Risk Score (PRS), to increase features matching[12, 54]

The ability to predict vulnerability and identify related predisposition biomarkers holds enormous possibilities including preventions tactics, treatments or simply avoidance. Equally important is the ability to identify resilience factors, those biomarkers or psychosocial "protective" characteristics, that can thwart or prevent the progress of alcohol dependence. Overall, our findings demonstrate the

importance of embedded ancestry, gender and age in the calculation of model prediction of the development of AUD. This approach we argue, should be expanded to any diagnosis or prediction of treatment response. We further show that the model based on various features from different realms (genetics, electrophysiology and FH) outperform prediction models based on singular-realm features. Wider selection of features with a narrower approach when choosing the sample will generate better prediction scores, enabling accurate anticipation of the development of an undesirable disorder. We also identified specific robust features of EEG and SNP measurement for each gender/ancestry group, further deepening our understanding of the predisposition of brain mechanisms underlying the future development of AUD. Future studies are required to further validate these results with larger cohorts, sampling uniformity and wider selection of features.

ACKNOWLEDGMENTS

The Collaborative Study on the Genetics of Alcoholism (COGA), Principal Investigators B Porjesz, V Hesselbrock, H Edenberg, L Bierut, includes 11 different centers: of Connecticut (V Hesselbrock); Indiana University (HJ Edenberg, J Nurnberger Jr, T Foroud); University of Iowa (S Kuperman, J

Kramer); SUNY Downstate (B Porjesz); Washington University in St Louis (L Bierut, J Rice, K Bucholz, A Agrawal); University of California at San Diego (M Schuckit); Rutgers University (J Tischfield, A Brooks); University of Texas Rio Grand Valley (L Almasy), Virginia Commonwealth University (D Dick), Icahn School of Medicine at Mount Sinai (A Goate), and Howard University (R Taylor). Other COGA collaborators include: L Bauer (University of Connecticut); J McClintick, L Wetherill, X Xuei, Y Liu, D. Lai, S O'Connor, M Plawecki, S Lourens (Indiana University); G Chan (University of Iowa; University of Connecticut); J Meyers, D Chorlian, C Kamarajan, A Pandey, J Zhang (SUNY Downstate); J-C Wang, M Kapoor, S Bertelsen (Icahn School of Medicine at Mount Sinai); A Anokhin, V McCutcheon, S Saccone (Washington University); J Salvatore, F Aliev, B Cho (Virginia Commonwealth University); and Mark Kos (University of Texas Rio Grand Valley). A Parsian and M Reilly are the NIAAA Staff Collaborators. We continue to be inspired by our memories of Henri Begleiter and Theodore Reich, founding PI and Co-PI of COGA, and also owe a debt of gratitude to other past organizers of COGA, including Ting-Kai Li, P Michael Conneally, Raymond Crowe and Wendy Reich, for their critical contributions. This national collaborative study is supported by an NIH Grant U10AA008401 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA). JLM is supported by

K01DA037914 from the National Institute on Drug Abuse (NIDA), JES acknowledges support from K01AA024152 (NIAAA) and AA acknowledges support from K02DA032573 (NIDA). Funding support for GWAS genotyping performed at the Johns Hopkins University Center for Inherited Disease Research was provided by the National Institute on Alcohol Abuse and Alcoholism, the NIH GEI (U01HG004438), and the NIH contract 'High throughput genotyping for studying the genetic contributions to human disease' (HHSN268200782096C). GWAS genotyping was also performed at the Genome Technology Access Center in the Department of Genetics at Washington University School of Medicine, which is partially supported by NCI Cancer Center Support Grant no. P30 CA91842 to the Siteman Cancer Center and by ICTS/CTSA Grant no. UL1RR024992 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Supplementary information is available at MP's website

References

1. Patrick ME, Schulenberg JE. Prevalence and predictors of adolescent alcohol use and binge drinking in the United States. *Alcohol Res* 2013; **35**(2)**:** 193-200.

2. Prescott CA, Kendler KS. Genetic and environmental contributions to alcohol abuse and dependence in a population-based sample of male twins. *Am J Psychiatry* 1999; **156**(1)**:** 34-40.

3. Whelan R, Watts R, Orr CA, Althoff RR, Artiges E, Banaschewski T *et al.* Neuropsychosocial profiles of current and future adolescent alcohol misusers. *Nature* 2014; **512**(7513)**:** 185-189.

4. Yang H, Liu J, Sui J, Pearlson G, Calhoun VD. A Hybrid Machine Learning Method for Fusing fMRI and Genetic Data: Combining both Improves Classification of Schizophrenia. *Front Hum Neurosci* 2010; **4:** 192.

5. Librenza-Garcia D, Kotzian BJ, Yang J, Mwangi B, Cao B, Pereira Lima LN *et al.* The impact of machine learning techniques in the study of bipolar disorder: A systematic review. *Neurosci Biobehav Rev* 2017; **80:** 538-554.

6. Sacchet MD, Prasad G, Foland-Ross LC, Thompson PM, Gotlib IH. Support vector machine classification of major depressive disorder using diffusion-weighted neuroimaging and graph theory. *Front Psychiatry* 2015; **6:** 21.

7. Bi J, Sun J, Wu Y, Tennen H, Armeli S. A machine learning approach to college drinking prediction and risk factor identification. *ACM Trans Intell Syst Technol* 2013; **4**(4)**:** 1-24.

8. Shim M, Hwang HJ, Kim DW, Lee SH, Im CH. Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level EEG features. *Schizophr Res* 2016; **176**(2-3)**:** 314-319.

9. Acharya UR, Sree SV, Chattopadhyay S, Suri JS. Automated diagnosis of normal and alcoholic EEG signals. *Int J Neural Syst* 2012; **22**(3)**:** 1250011.

10. Mumtaz W, Vuong PL, Xia LK, Malik AS, Bin Abd Rashidb R. Automatic diagnosis of alcohol use disorder using EEG features. *Knowl-Based Syst* 2016; **105:** 48-59.

11. Mumtaz W, Vuong P, Xia LK, Malik A, Bin Abd Rashid R. An EEG-based machine learning method to screen alcohol use disorder. *Cogn Neurodynamics* 2017; **11**(2)**:** 161-171.

12. Clarke TK, Adams MJ, Davies G, Howard DM, Hall LS, Padmanabhan S *et al.* Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=112 117). *Mol Psychiatry* 2017; **22**(10)**:** 1376-1384.

13. Jorgenson E, Thai KK, Hoffmann TJ, Sakoda LC, Kvale MN, Banda Y *et al.* Genetic contributors to variation in alcohol consumption vary by race/ethnicity in a large multi-ethnic genome-wide association study. *Mol Psychiatry* 2017; **22**(9)**:** 1359-1367.

14. Gelernter J, Kranzler HR, Sherva R, Almasy L, Koesterer R, Smith AH *et al.* Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. *Mol Psychiatry* 2014; **19**(1)**:** 41-49.

15. Polimanti R, Zhang H, Smith AH, Zhao H, Farrer LA, Kranzler HR *et al.* Genome-wide association study of body mass index in subjects with alcohol dependence. *Addict Biol* 2017; **22**(2)**:** 535-549.

16. Meyers JL, Zhang J, Wang JC, Su J, Kuo SI, Kapoor M *et al.* An endophenotype approach to the genetics of alcohol dependence: a genome wide association study of fast beta EEG in families of African ancestry. *Mol Psychiatry* 2017; **22**(12)**:** 1767-1775.

17. Pierce TW, Watson TD, King JS, Kelly SP, Pribram KH. Age differences in factor analysis of EEG. *Brain Topogr* 2003; **16**(1)**:** 19-27.

18. Zappasodi F, Marzetti L, Olejarczyk E, Tecchio F, Pizzella V. Age-Related Changes in Electroencephalographic Signal Complexity. *PLoS One* 2015; **10**(11)**:** e0141995.

19. Chorlian DB, Rangaswamy M, Manz N, Kamarajan C, Pandey AK, Edenberg H *et al.* Gender modulates the development of theta event related oscillations in adolescents and young adults. *Behav Brain Res* 2015; **292:** 342-352.

20. Ngun TC, Ghahramani N, Sanchez FJ, Bocklandt S, Vilain E. The genetics of sex differences in brain and behavior. *Front Neuroendocrinol* 2011; **32**(2)**:** 227-246.

21. Ali-Khan SE, Krakowski T, Tahir R, Daar AS. The use of race, ethnicity and ancestry in human genetic research. *Hugo J* 2011; **5**(1-4)**:** 47-63.

22. Sankar P, Cho MK. Genetics. Toward a new vocabulary of human genetic variation. *Science* 2002; **298**(5597)**:** 1337-1338.

23. Li L, Rakitsch B, Borgwardt K. ccSVM: correcting Support Vector Machines for confounding factors in biological data classification. *Bioinformatics* 2011; **27**(13)**:** i342-348.

24. Begleiter H, Porjesz B, Reich T, Edenberg HJ, Goate A, Blangero J *et al.* Quantitative trait loci analysis of human event-related brain potentials: P3 voltage. *Electroencephalogr Clin Neurophysiol* 1998; **108**(3)**:** 244-250.

25. Edenberg HJ, Bierut LJ, Boyce P, Cao M, Cawley S, Chiles R *et al.* Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14. *BMC Genet* 2005; **6 Suppl 1:** S2.

26. Reich T. A genomic survey of alcohol dependence and related phenotypes: results from the Collaborative Study on the Genetics of Alcoholism (COGA). *Alcohol Clin Exp Res* 1996; **20**(8 Suppl)**:** 133A-137A.

27. Giedd JN, Blumenthal J, Jeffries NO, Castellanos FX, Liu H, Zijdenbos A *et al.* Brain development during childhood and adolescence: a longitudinal MRI study. *Nat Neurosci* 1999; **2**(10)**:** 861-863.

28. Macleod S, Appleton RE. Neurological disorders presenting mainly in adolescence. *Arch Dis Child* 2007; **92**(2)**:** 170-175.

29. Klem GH, Luders HO, Jasper HH, Elger C. The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalogr Clin Neurophysiol Suppl* 1999; **52:** 3-6.

30. Oostenveld R, Fries P, Maris E, Schoffelen JM. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* 2011; **2011:** 156869.

31. Xu K, Kranzler HR, Sherva R, Sartor CE, Almasy L, Koesterer R *et al.* Genomewide Association Study for Maximum Number of Alcoholic Drinks in European Americans and African Americans. *Alcohol Clin Exp Res* 2015; **39**(7)**:** 1137-1147.

32. Wetherill L, Lai D, Johnson EC, Anokhin A, Bauer L, Bucholz KK *et al.* Genome-wide association study identifies loci associated with liability to alcohol and drug dependence that is associated with variability in reward-related ventral striatum activity in African- and European-Americans. *Genes Brain Behav* 2019; **18**(6)**:** e12580.

33. Lai D, Wetherill L, Bertelsen S, Carey CE, Kamarajan C, Kapoor M *et al.* Genome-wide association studies of alcohol dependence, DSM-IV criterion count and individual criteria. *Genes Brain Behav* 2019; **18**(6)**:** e12579.

34. Guyon I, Andr, #233, Elisseeff. An introduction to variable and feature selection. *J Mach Learn Res* 2003; **3:** 1157-1182.

35. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)* 1996; **58**(1)**:** 267-288.

36. Knight K, Fu W. Asymptotics for lasso-type estimators. 2000; **28**(5)**:** 1356-1378.

37. Ghosh D, Chinnaiyan AM. Classification and selection of biomarkers in genomic data using LASSO. *J Biomed Biotechnol* 2005; **2005**(2)**:** 147-154.

38. Shevade SK, Keerthi SS. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* 2003; **19**(17)**:** 2246-2253.

39. Rijsbergen CJv. *The Geometry of Information Retrieval*. Cambridge University Press2004.

40. Meunier D, Stamatakis EA, Tyler LK. Age-related functional reorganization, structural changes, and preserved cognition. *Neurobiol Aging* 2014; **35**(1)**:** 42-54.

41. Charles ST, Carstensen LL. Social and emotional aging. *Annu Rev Psychol* 2010; **61:** 383-409.

42. Newson JJ, Thiagarajan TC. EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies. *Front Hum Neurosci* 2018; **12:** 521.

43. van Kerkoerle T, Self MW, Dagnino B, Gariel-Mathis MA, Poort J, van der Togt C *et al.* Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proc Natl Acad Sci U S A* 2014; **111**(40)**:** 14332-14341.

44. Rangaswamy M, Porjesz B, Chorlian DB, Choi K, Jones KA, Wang K *et al.* Theta power in the EEG of alcoholics. *Alcohol Clin Exp Res* 2003; **27**(4)**:** 607-615.

45. Affan RO, Huang S, Cruz SM, Holcomb LA, Nguyen E, Marinkovic K. High-intensity binge drinking is associated with alterations in spontaneous neural oscillations in young adults. *Alcohol* 2018; **70:** 51-60.

46. Lopez-Caneda E, Cadaveira F, Correas A, Crego A, Maestu F, Rodriguez Holguin S. The Brain of Binge Drinkers at Rest: Alterations in Theta and Beta Oscillations in First-Year College Students with a Binge Drinking Pattern. *Front Behav Neurosci* 2017; **11:** 168.

47. Rangaswamy M, Porjesz B. From event-related potential to oscillations: genetic diathesis in brain (dys)function and alcohol dependence. *Alcohol Res Health* 2008; **31**(3)**:** 238-242.

48. Park SM, Lee JY, Kim YJ, Lee JY, Jung HY, Sohn BK *et al.* Neural connectivity in Internet gaming disorder and alcohol use disorder: A resting-state EEG coherence study. *Sci Rep* 2017; **7**(1)**:** 1333.

49. Peniston EG, Kulkosky PJ. Alpha-theta brainwave training and beta-endorphin levels in alcoholics. *Alcohol Clin Exp Res* 1989; **13**(2)**:** 271-279.

50. Kinreich S, Podlipsky I, Jamshy S, Intrator N, Hendler T. Neural dynamics necessary and sufficient for transition into pre-sleep induced by EEG neurofeedback. *Neuroimage* 2014; **97:** 19-28.

51. Lagopoulos J, Xu J, Rasmussen I, Vik A, Malhi GS, Eliassen CF *et al.* Increased theta and alpha EEG activity during nondirective meditation. *J Altern Complement Med* 2009; **15**(11)**:** 1187-1192.

52. Menon V, Uddin LQ. Saliency, switching, attention and control: a network model of insula function. *Brain Struct Funct* 2010; **214**(5-6)**:** 655-667.

53. Xue G, Lu Z, Levin IP, Bechara A. The impact of prior risk experiences on subsequent risky decision-making: the role of the insula. *Neuroimage* 2010; **50**(2)**:** 709-716.

54. Mies GW, Verweij KJH, Treur JL, Ligthart L, Fedko IO, Hottenga JJ *et al.* Polygenic risk for alcohol consumption and its association with alcohol-related phenotypes: Do stress and life satisfaction moderate these relationships? *Drug Alcohol Depend* 2018; **183:** 7-12.

Table 1: Selected significant models, classifying AUD and unaffected controls divided by ancestry, age and gender.

Model Features	Specificity (%)	Sensitivity (%)	Accuracy (%)	*AUC*	*F*
EEG					
EEG male	68.4	72.66	70.53	0.84	0.71
EEG female	73.5	77.86	75.68	0.86	0.76
EEG (12-15)	73.99	72.68	73.35	0.87	0.72
EA					
EEG + SNP	69.03	75.04	72.04	0.84	0.73
EEG + SNP Female	76.82	81	78.91	0.9	0.79
EEG +SNP Male mother AUD	70.94	68.08	69.54	0.78	0.69
EEG +SNP Female father AUD	70.15	78.22	74.18	0.92	0.76
EEG +SNP Female mother AUD	75.16	79.48	77.32	0.89	0.78
EEG+SNP (12-15)	68.43	79.23	74.2	0.89	0.76
AA					
EEG + SNP	74.14	78.43	76.29	0.9	0.77
EEG + SNP Female	71.02	87.67	79.33	0.99	0.81
EEG +SNP Male mother AUD	74.69	67.78	71.23	0.83	0.71
EEG +SNP Female father AUD	85.71	89.38	87.55	0.99	0.89
EEG +SNP Female mother AUD	81.3	92.92	87.11	0.99	0.88
EEG+SNP (12-15)	79.55	79.52	79.54	0.93	0.79
EEG+SNP (16-19)	65.46	85.56	76.53	0.92	0.80
EEG+SNP (20-30)	73.2	73.18	73.19	0.97	0.71

Note: F tests were used for comparisons between the two groups. *AUC (Area Under the Curve*) calculations were used for classification analysis in order to determine which of the used models predicts the labels best. Values are means of the 10 CV fold model calculation.

Legends

Figure 1. A data-flow diagram.

Figure 2. Model accuracy by ancestry. Classification obtained by the only EEG features, only SNP features and by the combined EEG and SNP features for EA and AA samples. Results indicate that the combined model has higher accuracy than the EEG based model, and the SNP based model. The error bars are standard deviations. $*p < .05$, $**p < .01$.

Figure 3. Model accuracy by gender and ancestry. Classification accuracy obtained by the only EEG features, only SNP features and by the combined EEG and SNP features, stratified by gender. Results indicate higher accuracy scores for the female compared to male in both EA and AA samples for the three models-based features. The error bars are standard deviations. $*p < .05$, $**p < .01$.

Figure 4. Model accuracy by gender, family history and ancestry. Mother AUD and father AUD features were added to the female and male models. Results highlight ancestry and gender differences of the effect of parent AUD over the accuracy of the model. For both AA and EA, male and female samples, mother AUD feature increased model accuracy. Father AUD increased the accuracy of the combined model for the AA female sample. The error bars are standard deviations. $*p < .05$, $**p < .01$.

Fig 1.

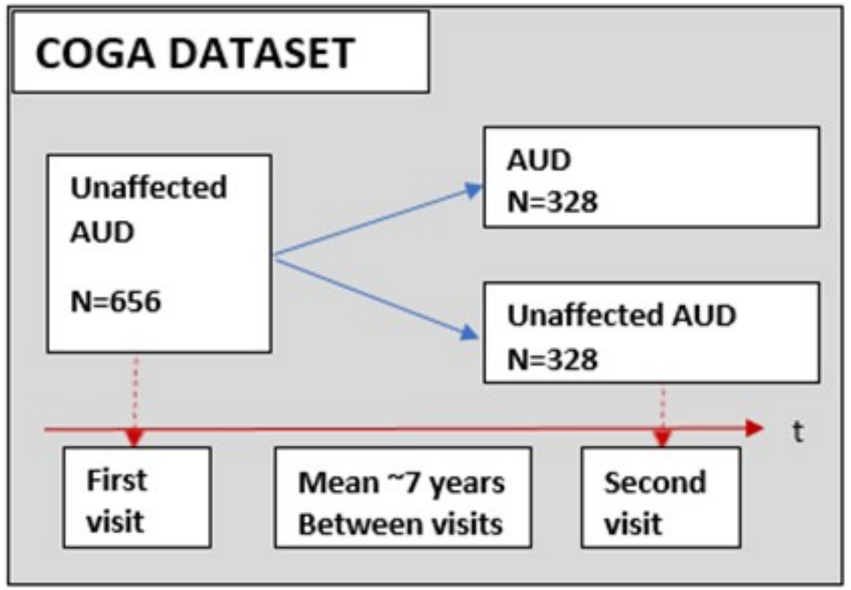

COGA DATASET
Unaffected AUD
N=656
AUD
N=328
Unaffected AUD
N=328
t
First visit
Mean ~7 years Between visits
Second visit

Fig 2.

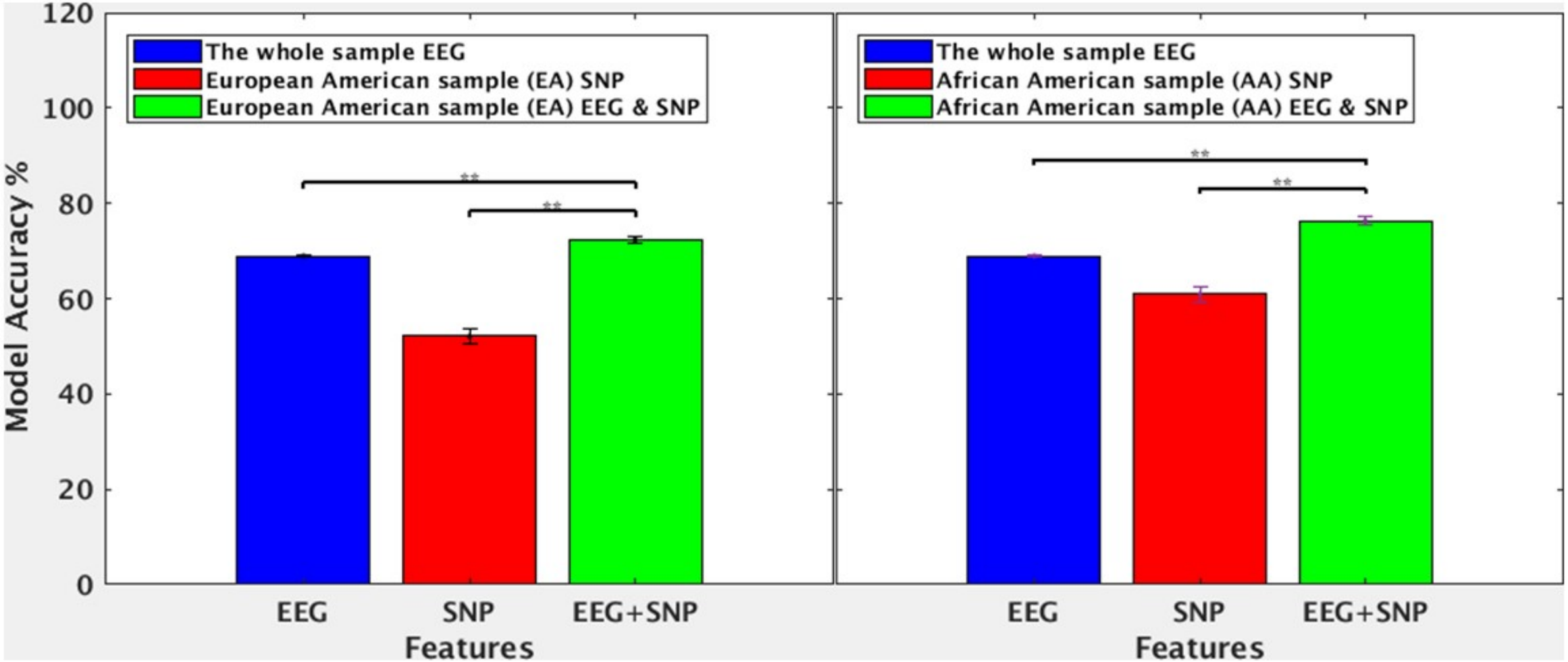

120
100
80
60
40
20
0
Model Accuracy %
The whole sample EEG
European American sample (EA) SNP
European American sample (EA) EEG & SNP
**
**
EEG
SNP
EEG+SNP
Features
The whole sample EEG
African American sample (AA) SNP
African American sample (AA) EEG & SNP
**
**
EEG
SNP
EEG+SNP
Features

Fig 3.

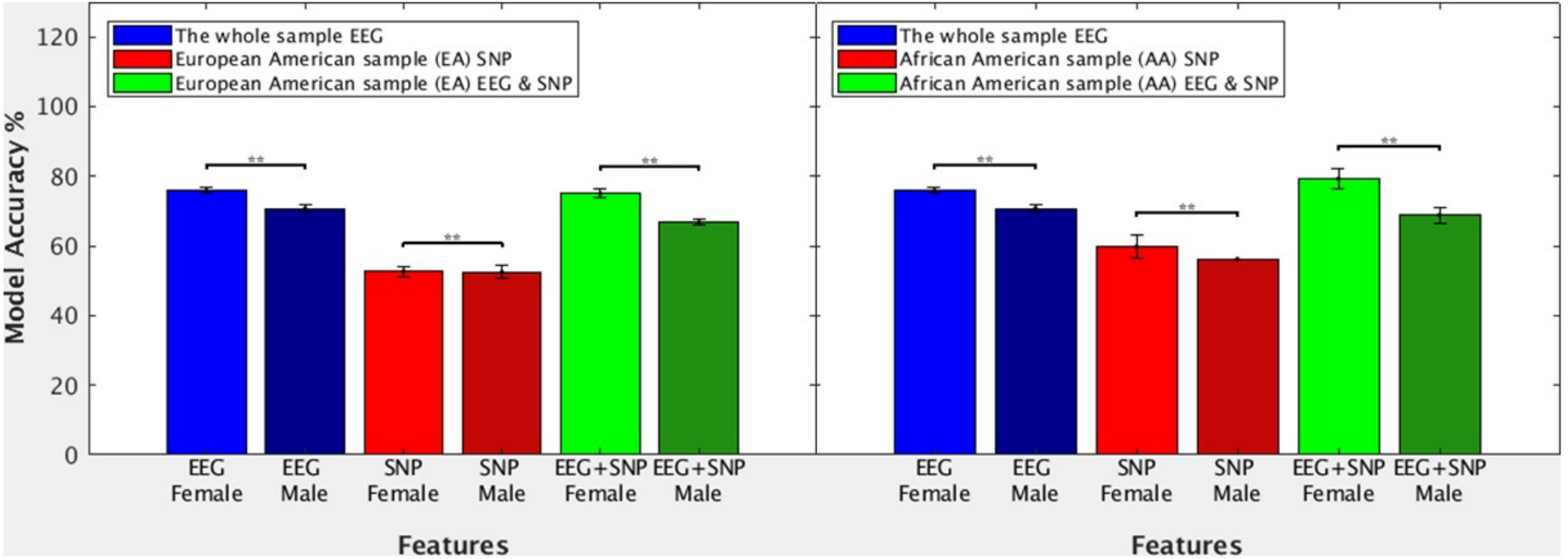
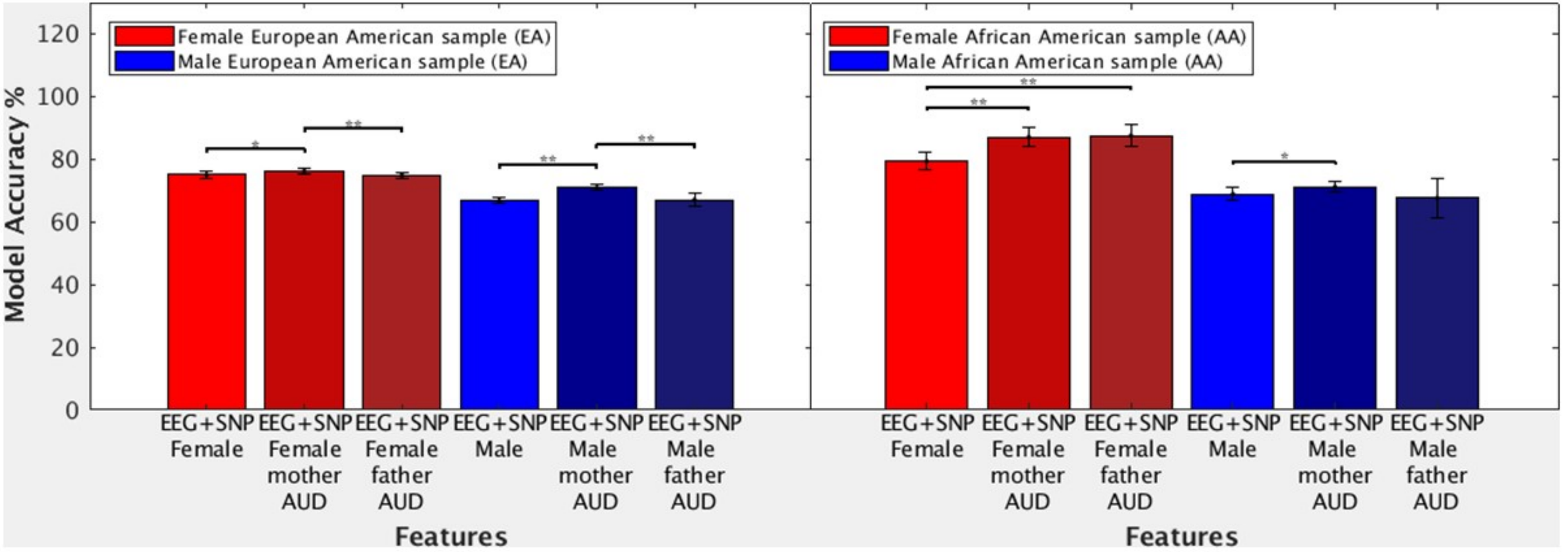

The whole sample EEG
European American sample (EA) SNP
European American sample (EA) EEG & SNP
Model Accuracy %
120
100
80
60
40
20
0
**
**
**
EEG Female
EEG Male
SNP Female
SNP Male
EEG+SNP Female
EEG+SNP Male
Features
The whole sample EEG
African American sample (AA) SNP
African American sample (AA) EEG & SNP
**
**
**
EEG Female
EEG Male
SNP Female
SNP Male
EEG+SNP Female
EEG+SNP Male
Features

Fig 4.


Female European American sample (EA)
Male European American sample (EA)
Model Accuracy %
120
100
80
60
40
20
0
*
**
**
**
EEG+SNP Female
EEG+SNP Female mother AUD
EEG+SNP Female father AUD
EEG+SNP Male
EEG+SNP Male mother AUD
EEG+SNP Male father AUD
Features
Female African American sample (AA)
Male African American sample (AA)
**
**
*
EEG+SNP Female
EEG+SNP Female mother AUD
EEG+SNP Female father AUD
EEG+SNP Male
EEG+SNP Male mother AUD
EEG+SNP Male father AUD
Features